Dickinson College

Exploring the factors influencing career change decisions: A quantitative analysis of motivational drivers, including career development opportunities and current employment & living conditions.

Amanda Tran

DATA 300: Statistical & Machine Learning
Professor Xiexin Liu
May 11, 2023

Introductions:

In numerous discussions held with industry experts, the subject of career transitions has been a recurring topic. Career specialists tend to bring up their transition from one career, or company, to another within the overall sentiment of a general desire to find a better fit job-wise. More in-depth investigation into career transitions also highlight the significance of training, company prospects, and other pertinent factors in shaping an individual's career trajectory. According to Carless & Arnup (2011), a variety of factors including age, gender and education level can drive one's decision to change their career path. Vaitenas & Wiener (1977), on another hand, referred to motivation as an open field of inquiry in its drive towards career changes.

Objectives:

This research project aims to investigate a dataset of individuals who have received training from a prominent Big Data firm and have demonstrated an interest in working for the company. The dataset presents an opportunity to (1) identify key metrics that contribute to an individual's decision to leave their previous job and (2) identify the most suitable candidates for employment within the company. It is key that these metrics are discovered so that this company can discover the best approach to implement their resources, specifically in the career development and recruiting programs.

Overview of analysis subjects:

For the purpose of analysis, a publicized dataset was compiled by an anonymous Big Data company after their career development seminar. The dataset originally contained 19200 observations of 14 variables, 2 of which are numerical and 12 of which are categorial variables. The list of variables and explanations for each variable follows:

Table 1:

List of Variables (Response Variable in bold)

Categorical Variables:	Numerical Variables:
• enrollee_id: Unique ID for enrollee • city: City code • gender: The gender of the candidate • relevant_experience: Whether the candidate had relevant experience • enrolled_university: Type of University course enrolled, if any • education_level: The candidate's education level • major_discipline: The candidate's major/discipline • experience: The candidate's total experience in years • company_size: Number of employees in the candidate's current company • company_type: Type of the candidate's current employer • last_new_job: Difference in years between previous job and current job • **target: 0 – Candidate not looking for job change, 1 – Candidate looking for a job change**	• city_development_index: Developement index of the city (scaled) • training_hours: Training hours completed

Methodology:

a. Data Pre-Processing:

As there exist many categorical variables and unaddressed NA values in this dataset, some cleaning has been done to improve the analysis of this dataset:

(1) Several categorical variables have NULL observations that were grouped with the category "Other" or a similar neutral representation (e.g., "no_enrollment" in "enrolled_university"). These variables are "gender", "enrolled_university", "major_discipline" and "company_type".

(2) Categorical variables like "experience", "last_new_job" and "company_size" have too many categories, which may confuse statistical models. The most feasible approach for handling these was to group them into bigger categories: (a) "experience" was re-arranged into 3 categories: 10 years, 10-15 years, >15 years, (b) "company_size" was re-arranged into 3 categories: Small (<100), Midsize (<5000) and Large (5000+), (c) "last_new_job" into 2 binary values: 0 (no past experience) and 1 (>1 years in past experiences).

(3) The remaining NAs are omitted

(4) Finally, categorical variables are converted to dummy binary variables.

After data cleaning is completed, irrelevant variables that might distort statistical learning are omitted: "enrollee_id" and "city". There remain 10809 observations of 28 variables.

b. Classification Analysis:

As proposed, this project aims to investigate relationships between different metrics and whether someone wants to switch careers. Thus, the variable "target" will be the response variable in a classification model. Due to the process of data cleaning, there is ample room for imbalance in class distributions within categorical variables. Therefore, **resampling**, specifically over-, and under-sampling was conducted on the dataset before any model optimization to save on resources and time for the business operation. The optimization process was followed by **variable selection methods** (i.e., best subset selection, ridge regression) to set preliminary expectations of variables that showed most association to a candidate's career targets. While **tree-based methods** are implemented

to present solutions with better interpretability and identify deterministic factors, **support vector machines** are also involved to create an alternative with higher flexibility in predicting future data.

Results:

a. Logistic Regression Analysis:

(1) The traditional, full logistic regression model consists of all 28 variables and returned an AIC score of 8850.6. At a 0.5 threshold, the model's accuracy is 0.8205365.

(2) At a 0.5 threshold, the model based on over-sampling and under-sampling returned an accuracy rate of 0.8159112. Considering the implementation of tree-based methods in latter analyses, we moved forward with an under-sampled subset of the original data, in spite of the minimal decrease in accuracy.

(3) Based on the progress from resampling, further variable selection methods were used on the classification model. Two approaches were tested: a naïve approach using best subset selection and a ridge approach.

   a. Via best subset selection, it was recommended that we employ 5 variables into our model: "Edu_Level_High_School", "city_development_index", "relevant_experience", "company_size_Mid_size" and "company_type_Public_Sector".

   b. Via optimal ridge regression, the best lambda ridge for our model was recommended to be 0.03997048.

Running predictions on the "validation" set, we have graphs of the ROC curve:

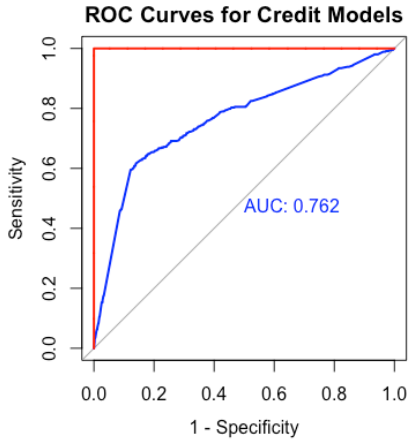

Figure 1: ROC Curves for Logistic Regression

At a 0.5 threshold, it seems that the ridge regression model achieves the ultimate accuracy of 1

b. Decision Trees and Tree-based Methods:

A preliminary examination of the dataset showed the variable "city_development_index" being a very strong variable in splitting the data that overshadowed the decision-making power of other variables. Thus, for the sake of conducting tree-based analysis, the dataset had been under-sampled down to 2174 observations/class (total of 4348 observations), and the aforementioned category was excluded from analysis models.

A traditional classification tree was first applied to our dataset. Without the "city_development_index", the tree produced showed that: a candidate with (1) a STEM major who has had (2) less than 10 years of experience in their current job is most likely to commit to a switch in career to the Data company that provided this training (fig. 2). Since there was some overlapping area in the "experience" category due to the variable coding, a pruned tree of 2 nodes was implemented to reduce unnecessary complexity. Based on this tree, having a STEM major is the sole impactful determinant of whether a candidate is likely to switch career post-training (fig. 3).

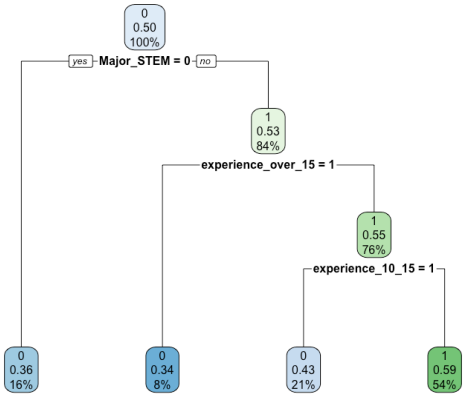

Figure 2: Traditional Classification Tree (4 Terminal Nodes)

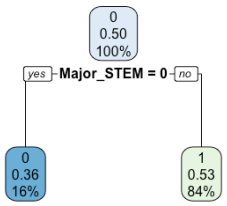

Figure 3: Pruned Classification Tree (2 Terminal Nodes)

With a sense of what the deterministic "predictors" could be, further tree-based methods were carried out. Bagging at B = 500 trees of 26 predictors, and random forest with m = 5 predictors were carried out in hopes of reducing variance. Interestingly, the results from both methods suggested that training hours, experience and college major outstandingly contributes to a decrease in GINI index, a measurement of misclassification, among both tweaked tree models (fig. 4 & fig. 5).

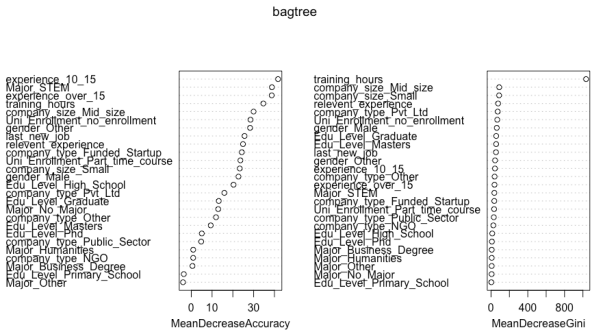

Figure 4: Importance of variables in bagged classification tree, ranked in contribution to decreasing GINI index and improving accuracy

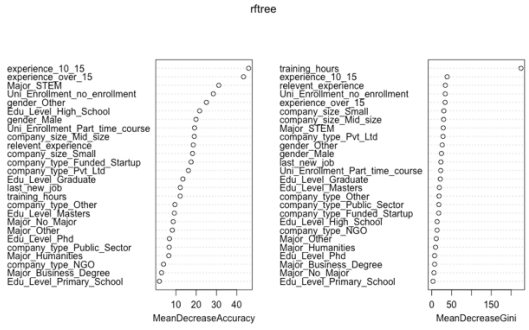

Figure 5: Importance of variables in random-forest classification tree, ranked in contribution to decreasing GINI index and improving accuracy

Boosting the existing decision tree reaffirmed this finding on training hours. A 5-fold cross-validation within the grid of n trees (n) = [100,1000] and interaction depth (d) = [1,5] returned the optimal parameters of n = 1000 and d= 1; with 1000 trees, each built sequentially upon the residuals of its preceding trees, the results drew attention to (1) a major in STEM, (2) prior experience of less than 10 years, (3) training hours spent, and (4) undeclared/non-binary gender.

In order to best study the output of this training program, it is key that the company correctly identifies the candidates most likely to demonstrate interest and commitment post-training to best allocate their communications effort; thus, the predictive powers of all decision trees carried out shall be assessed using the True Positive Rate (Sensitivity/Recall):

Table 2: Sensitivity/Recall Rate among Tree-based methods

Sensitivity/Recall	Unpruned Tree	Pruned Tree	Bagging	Random Forest	Boosting
	0.4770	0.4355	0.5899	0.5300	0.5579

c. Support Vector Machine:

A Support Vector Machine (SVM) with a linear kernel of cost (i.e., the trade-off penalty between training error and validation error) 1 was generated. The model found 2025 "support vectors", which were the data points closest to the decision boundary that separates the two classes; 967 of them were from class 0 (i.e., not looking for a job change) and the other 1058 belong to class 1. The predictive specificity of this model was 74.78%.

Discussion and Future Interpretations:

Our overall classification models' results', though vary in approaches, all led to intriguing results. Based on the optimized logistic regression result through best subset selection and Ridge regression, an individual's inclination to change jobs can be strongly associated with:

- An education level beyond high school
- Current living situation, based on city development index
- Having prior/current experience in a mid-sized (<5000) or public sector organization.

On the other hand, our tree-based models offer a more interpretable guideline of the deterministic variables contributing to a career change. Candidates who are highly likely to seek a career switch after the training program belong to including, but not limited to, these demographics:

- Candidates pursuing or having completed a degree in a STEM field.
- Candidates who are in entry-level/associate roles/stages of their career.
- Candidates who are frequent applicants, or who demonstrated enthusiasm and commitment to multiple career development programs presented by the company.
- Candidates who are currently located in middle-to-lesser-developed cities.

Finally, the Support Vector Machine implemented offer a more flexible model with increased accuracy, which supports the Big Data company's long-term implementation of these metrics in predicting the output of their training programs.

The findings from this quantitative analysis supports prior research conducted that cover the similar topic of motivation behind career switches such as education (Carless & Arnup, 2011); yet, these findings also offer additional evidence on the influence of living and working conditions on changing jobs, which furthers the discourse on self-motivation and wellness in the workplace and academic research (Vaitenas & Wiener, 1977). By defining its ideal training demographics, this Big Data company will save on time and resources choosing candidates for programs that serve as pipelines for employment in the future. Based on the given results from decision-tree-based analysis, this company and its Human Resources Department can narrow down their demographic for recruitment of future training/bootcamp-style career development.

Bibliography

Carless, S. A., & Arnup, J. L. (2011). A longitudinal study of the determinants and outcomes of career change. *Journal of vocational behavior*, *78*(1), 80-91.

Vaitenas, R., & Wiener, Y. (1977). Developmental, emotional, and interest factors in voluntary mid-career change. *Journal of Vocational Behavior*, *11*(3), 291-3